

Stochastic optimization with decision-dependent distributions

Dmitriy Drusvyatskiy
Mathematics, University of Washington

Joint work with L. Xiao (Facebook AI)

IFDS 2020

What this talk is about.

Stochastic optimization with state-dependent distributions

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(x)} [\ell(x, z)] + r(x)$$

What this talk is about.

Stochastic optimization with state-dependent distributions

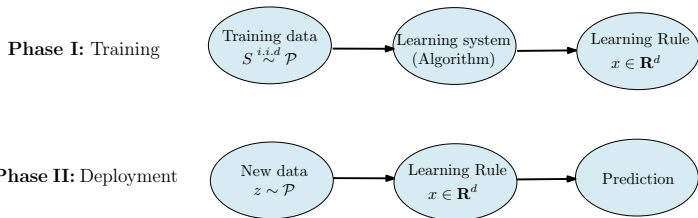
$$\min_x \mathbb{E}_{z \sim \mathcal{D}(x)} [\ell(x, z)] + r(x)$$

Building on framework of [Perdomo-Zrnic-Dünner-Hardt](#):

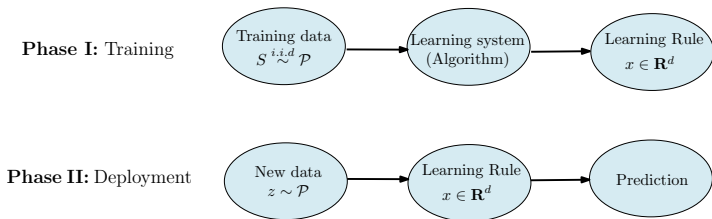
- ▶ “Performative prediction” (ICML 2020)
- ▶ “Stochastic optimization for performative prediction” (NeurIPS 2020)

Introduction

Pipeline of Supervised Learning

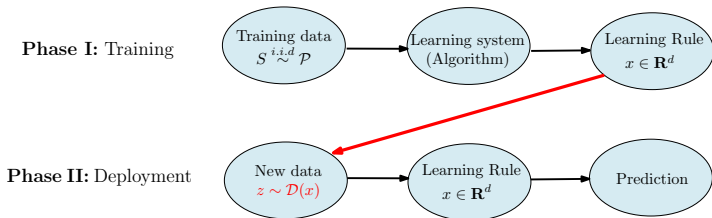


Pipeline of Supervised Learning

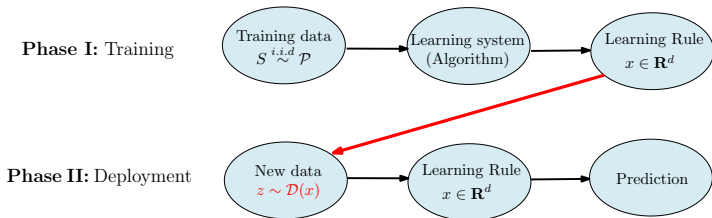


Key Assumption: Both test data and training data drawn from $\mathcal{P}$

Learning systems do not exist in isolation...



Learning systems do not exist in isolation...



Example (passive interaction):

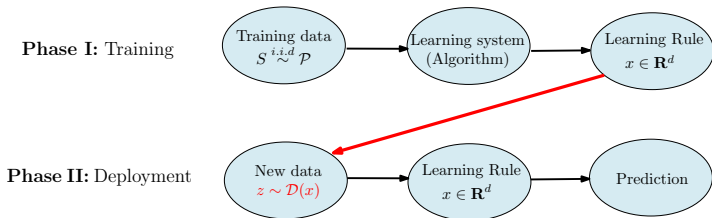
Bank loan approval influences debt/credit score/#loans.

Example (active interaction):

[strategic behavior/gaming]

Individuals alter features to increase likelihood of loan approval.

Learning systems do not exist in isolation...



Example (passive interaction):

Bank loan approval influences debt/credit score/#loans.

Example (active interaction): [strategic behavior/gaming]

Individuals alter features to increase likelihood of loan approval.

Perdomo-Zrnich-Dünner-Hardt '20 call this setting **performative prediction**

Optimization model

Stochastic optimization with state-dependent distributions

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(x)} [\ell(x, z)] + r(x)$$

where

- $\mathcal{D}(x)$ are **state-dependent** distributions accessible by sampling
- $\ell(\cdot, z)$ is a convex loss
- $r(\cdot)$ is convex structure-inducing regularizer

Optimization model

Stochastic optimization with state-dependent distributions

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(x)} [\ell(x, z)] + r(x)$$

where

- $\mathcal{D}(x)$ are **state-dependent** distributions accessible by sampling
- $\ell(\cdot, z)$ is a convex loss
- $r(\cdot)$ is convex structure-inducing regularizer

Decision x is judged according to $\mathcal{D}(x)$.

Optimization model

Stochastic optimization with state-dependent distributions

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(x)} [\ell(x, z)] + r(x)$$

where

- $\mathcal{D}(x)$ are **state-dependent** distributions accessible by sampling
- $\ell(\cdot, z)$ is a convex loss
- $r(\cdot)$ is convex structure-inducing regularizer

Decision x is judged according to $\mathcal{D}(x)$.

Bad news: nonsmooth, nonconvex

Two paths forward:

1. Impose “smoothness” or “structure” on $\mathcal{D}(\cdot)$ and solve.
e.g. Ahmed '00, Dupačová '06, Goel-Grossman '06, Hassani et al. '20
2. Settle for a related and efficiently computable solution concept.
Perdomo-Zrnic-Dünner-Hardt '20

Equilibrium

Notation:

$$f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \nabla f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \nabla \ell(x, z)$$

Equilibrium

Notation:

$$f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \nabla f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \nabla \ell(x, z)$$

Definition (Perdomo et al '20)

A point $\bar{x}$ is at **equilibrium** for $\mathcal{D}(\cdot)$ if

$$\bar{x} = \operatorname{argmin}_x \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} \ell(x, z) + r(x)$$

“No incentive to alter $\bar{x}$ based only on response $\mathcal{D}(\bar{x})$.”

Equilibrium

Notation:

$$f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \nabla f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \nabla \ell(x, z)$$

Definition (Perdomo et al '20)

A point $\bar{x}$ is at **equilibrium** for $\mathcal{D}(\cdot)$ if

$$\bar{x} = \operatorname{argmin}_x f_{\bar{x}}(x) + r(x)$$

“No incentive to alter $\bar{x}$ based only on response $\mathcal{D}(\bar{x})$.”

Equilibrium

Notation:

$$f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \nabla f_y(x) = \mathbb{E}_{z \sim \mathcal{D}(y)} \nabla \ell(x, z)$$

Definition (Perdomo et al '20)

A point $\bar{x}$ is at **equilibrium** for $\mathcal{D}(\cdot)$ if

$$\bar{x} = \operatorname{argmin}_x f_{\bar{x}}(x) + r(x)$$

“No incentive to alter $\bar{x}$ based only on response $\mathcal{D}(\bar{x})$.”

Algorithmically: these are fixed points of the map

$$S(y) := \operatorname{argmin}_x f_y(x) + r(x).$$

$\Rightarrow$ suggests a fixed-point algorithm

Performative prediction

Repeated minimization:

$$x_{t+1} = \operatorname{argmin}_x \mathbb{E}_{z \sim \mathcal{D}(x_t)} [\ell(x, z)] + r(x)$$

Performative prediction

Repeated minimization:

$$x_{t+1} = \operatorname{argmin}_x \mathbb{E}_{z \sim \mathcal{D}(x_t)} [\ell(x, z)] + r(x)$$

Algorithms for static problems **heuristically** generalize.

Example: Proximal stochastic gradient

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \operatorname{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Similar for dual averaging, prox-point, clipped gradient, fast gradient, ...

Performative prediction

Repeated minimization:

$$x_{t+1} = \operatorname{argmin}_x \mathbb{E}_{z \sim \mathcal{D}(x_t)} [\ell(x, z)] + r(x)$$

Algorithms for static problems **heuristically** generalize.

Example: Proximal stochastic gradient

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \operatorname{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Similar for dual averaging, prox-point, clipped gradient, fast gradient, ...

Perdomo et al. '20:

1. Proposed this framework
2. Existence of equilibria
3. Convergence of repeated minimization
4. Convergence of (stochastic) projected gradient method

Our contribution

Meta Thm: *Algorithms that sample according to $\mathcal{D}(x_t)$ can be viewed as the same algorithms applied to the **static problem***

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} [\ell(x, z)] + r(x)$$

where “bias” $\rightarrow 0$ linearly as $x_t \rightarrow \bar{x}$.

Our contribution

Meta Thm: *Algorithms that sample according to $\mathcal{D}(x_t)$ can be viewed as the same algorithms applied to the **static problem***

$$\min_x \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} [\ell(x, z)] + r(x)$$

where “bias” $\rightarrow 0$ linearly as $x_t \rightarrow \bar{x}$.

Recipe:

algorithms for static problems $\hookrightarrow$ “mildly dynamic”

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

$$\boxed{\nabla f_x(x) \quad \text{and} \quad \nabla f_{\bar{x}}(x)}$$

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

$$\nabla f_x(x) \quad \text{and} \quad \nabla f_{\bar{x}}(x)$$

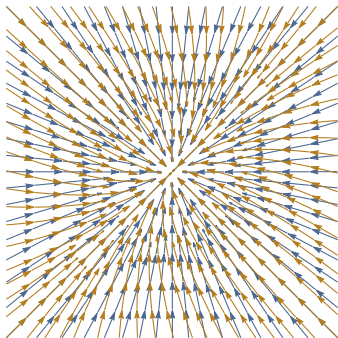


Figure: $\rho = 0.25$

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

$$\nabla f_x(x) \quad \text{and} \quad \nabla f_{\bar{x}}(x)$$

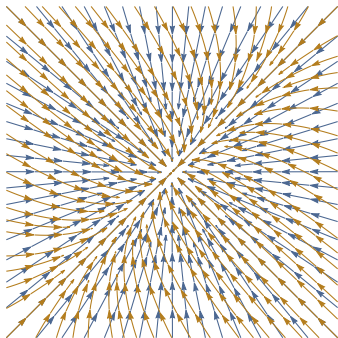


Figure: $\rho = 0.5$

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

$$\nabla f_x(x) \quad \text{and} \quad \nabla f_{\bar{x}}(x)$$

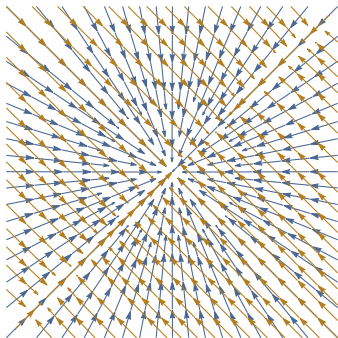


Figure: $\rho = 0.99$

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

$$\nabla f_x(x) \quad \text{and} \quad \nabla f_{\bar{x}}(x)$$

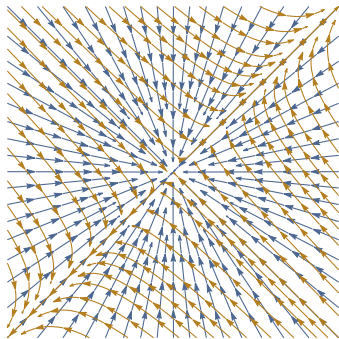


Figure: $\rho = 1.25$

Numerical illustration

Chasing the mean:

$$\min_{x \in \mathbb{R}^2} \mathbb{E}_{z \sim \mathcal{D}(x)} \|x - z\|^2 \quad \text{where} \quad \mathcal{D}(x_1, x_2) = N(\rho(x_2, x_1), I)$$

Equilibrium point $\bar{x} = (0, 0)$.

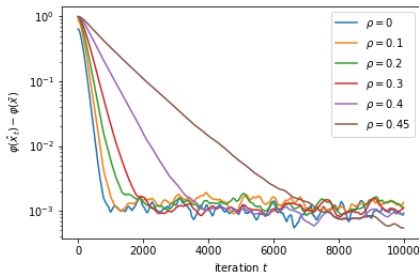


Figure: Stochastic gradient method (fixed $\eta > 0$)

Conclusion: meta-theorem seems valid when $\rho \in (0, 1)$

Outline

- ▶ Notation and assumptions
- ▶ Two deviation inequalities
- ▶ Reduction to online convex optimization
- ▶ Stochastic (accelerated) gradient
- ▶ Model-based algorithms

Notation:

- ▶ Fix $\ell: \mathbb{R}^d \times Z \rightarrow \mathbb{R}$ where Z is a metric space
- ▶ $\mathbb{P} = \{\text{probability measures on } Z\}$ with Wasserstein-1 distance $W_1(\mu, \nu)$

Assumption:

- ▶ **(smoothness/convexity)** Loss $\ell(\cdot, z)$ is α -strongly convex and

$$\|\nabla \ell(x, z) - \nabla \ell(x, z')\| \leq \beta \cdot d(z, z')$$

$$\|\nabla \ell(x, z) - \nabla \ell(x', z)\| \leq L \cdot \|x - x'\|$$

- ▶ **(sensitivity)** It holds:

$$W_1(\mathcal{D}(x), \mathcal{D}(y)) \leq \gamma \cdot \|x - y\|$$

Conditioning measures:

$$\kappa = \frac{L}{\alpha} \quad \text{and} \quad \rho = \frac{\gamma\beta}{\alpha}$$

Interesting regime is $\rho \in (0, 1)$

Recall **repeated minimization**:

$$x_{t+1} = \operatorname{argmin}_x f_{x_t}(x) + r(x)$$

Theorem (Perdomo et al. '20)

If $\rho < 1$, then repeated minimization converges to $\bar{x}$ at linear rate ρ .

If $\rho > 1$, then repeated minimization may diverge.

Interesting regime is $\rho \in (0, 1)$

Recall **repeated minimization**:

$$x_{t+1} = \operatorname{argmin}_x f_{x_t}(x) + r(x)$$

Theorem (Perdomo et al. '20)

If $\rho < 1$, then repeated minimization converges to $\bar{x}$ at linear rate ρ .

If $\rho > 1$, then repeated minimization may diverge.

True for wider class of algorithms including **proximal point method**

$$x_{t+1} = \operatorname{argmin}_x f_{x_t}(x) + r(x) + \frac{1}{2\eta} \|x - x_t\|^2$$

Theorem (D-Xiao '20)

*If $\rho < 1$, then **prox-point method** converges to $\bar{x}$ at linear rate $1 - \frac{1-\rho}{1+(\alpha\eta)^{-1}}$.*

Interesting regime is $\rho \in (0, 1)$

Recall **repeated minimization**:

$$x_{t+1} = \operatorname{argmin}_x f_{x_t}(x) + r(x)$$

Theorem (Perdomo et al. '20)

If $\rho < 1$, then repeated minimization converges to $\bar{x}$ at linear rate ρ .

If $\rho > 1$, then repeated minimization may diverge.

True for wider class of algorithms including **proximal point method**

$$x_{t+1} = \operatorname{argmin}_x f_{x_t}(x) + r(x) + \frac{1}{2\eta} \|x - x_t\|^2$$

Theorem (D-Xiao '20)

*If $\rho < 1$, then **prox-point method** converges to $\bar{x}$ at linear rate $1 - \frac{1-\rho}{1+(\alpha\eta)^{-1}}$.*

Advantage: prox-point is always “distributionally stable”

Regularization experimentally helps!

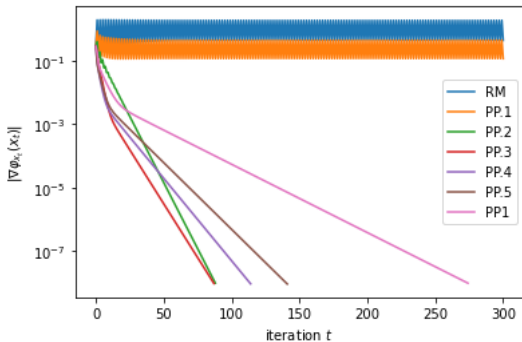


Figure: Strategic classification with $\rho > 1$.

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Question: how do ∇f_y and $\mathcal{G}_y$ vary with y ?

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Question: how do ∇f_y and $\mathcal{G}_y$ vary with y ?

Lemma (Gradient deviation)

For all $x, y \in \mathbb{R}^d$ it holds:

$$\sup_{x \in \mathbb{R}^d} \|\nabla f_{\mathbf{y}}(x) - \nabla f_{\mathbf{y}'}(x)\| \leq \gamma\beta \cdot \|\mathbf{y} - \mathbf{y}'\|$$

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Question: how do ∇f_y and $\mathcal{G}_y$ vary with y ?

Lemma (Gradient deviation)

For all $x, y \in \mathbb{R}^d$ it holds:

$$\sup_{x \in \mathbb{R}^d} \|\nabla f_{\mathbf{y}}(x) - \nabla f_{\mathbf{y}'}(x)\| \leq \gamma\beta \cdot \|\mathbf{y} - \mathbf{y}'\|$$

Implication: $\text{Bias}(x) := \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\| \leq \gamma\beta \cdot \|x - \bar{x}\|.$

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Question: how do ∇f_y and $\mathcal{G}_y$ vary with y ?

Lemma (Gradient deviation)

For all $x, y \in \mathbb{R}^d$ it holds:

$$\sup_{x \in \mathbb{R}^d} \|\nabla f_y(x) - \nabla f_{y'}(x)\| \leq \gamma\beta \cdot \|y - y'\|$$

Implication: $\text{Bias}(x) := \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\| \leq \gamma\beta \cdot \|x - \bar{x}\|.$

Lemma (Gap deviation)

All $x, x' \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}^d$ satisfy:

$$\mathcal{G}_y(x, x') - \mathcal{G}_{y'}(x, x') \leq \gamma\beta \cdot \|x - x'\| \cdot \|y - y'\|$$

Two deviation inequalities

Define

$$f_y(x) := \mathbb{E}_{z \sim \mathcal{D}(y)} \ell(x, z) \quad \text{and} \quad \mathcal{G}_y(x, x') := f_y(x) - f_y(x').$$

Question: how do ∇f_y and $\mathcal{G}_y$ vary with y ?

Lemma (Gradient deviation)

For all $x, y \in \mathbb{R}^d$ it holds:

$$\sup_{x \in \mathbb{R}^d} \|\nabla f_y(x) - \nabla f_{y'}(x)\| \leq \gamma\beta \cdot \|y - y'\|$$

Implication: $\text{Bias}(x) := \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\| \leq \gamma\beta \cdot \|x - \bar{x}\|.$

Lemma (Gap deviation)

All $x, x' \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}^d$ satisfy:

$$\mathcal{G}_y(x, x') - \mathcal{G}_{y'}(x, x') \leq \gamma\beta \cdot \|x - x'\| \cdot \|y - y'\|$$

Implication: $\mathcal{G}_x(x, \bar{x}) - \mathcal{G}_{\bar{x}}(x, \bar{x}) \leq \gamma\beta \cdot \|x - \bar{x}\|^2$ offset by strong convexity

Reduction to online convex optimization

Online convex optimization is a repeated game:

Round $t \geq 1$:

- ▶ Player chooses $x_t \in \text{dom } r$
- ▶ Nature reveals function ℓ_t and player pays $\ell_t(x_t)$

Player's goal: Minimize the regret

$$R_t := \sum_{i=1}^t (\ell_i(x_i) + r(x_i)) - \min_x \sum_{i=1}^t (\ell_i(x) + r(x)),$$

Algorithms: prox-grad. (Duchi-Singer '09), dual averaging (Xiao '10),
Follow-The-Regularized-Leader (FTRL) (McMahan '11)

Guarantees:

$$\left\{ \begin{array}{l} \ell_t \text{ are } \alpha\text{-strongly convex on } \text{dom } r \\ \ell_t \text{ are } G\text{-Lipschitz on } \text{dom } r \end{array} \right\} \implies R_t = \mathcal{O} \left(\frac{G^2 \log t}{\alpha} \right)$$

Reduction to online convex optimization

Recall the **equilibrium problem**:

$$\min_x \varphi(x) := f_{\bar{x}}(x) + r(x) \quad \text{where} \quad f_{\bar{x}}(x) = \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} [\ell(x, z)].$$

Reduction to online convex optimization

Recall the **equilibrium problem**:

$$\min_x \varphi(x) := f_{\bar{x}}(x) + r(x) \quad \text{where} \quad f_{\bar{x}}(x) = \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} [\ell(x, z)].$$

Theorem (D-Xiao '20)

Suppose $\rho \in (0, \frac{1}{2})$. Run an online algorithm where in iteration t , nature draws $z_t \sim \mathcal{D}(x_t)$ and declares $\ell_t(x_t) = \ell(x_t, z_t)$. Then

$$\mathbb{E} \left[\varphi \left(\frac{1}{t} \sum_{i=1}^t x_i \right) - \varphi(\bar{x}) \right] \leq \frac{\mathbb{E}[R_t]}{(1 - 2\rho)t},$$

Reduction to online convex optimization

Recall the **equilibrium problem**:

$$\min_x \varphi(x) := f_{\bar{x}}(x) + r(x) \quad \text{where} \quad f_{\bar{x}}(x) = \mathbb{E}_{z \sim \mathcal{D}(\bar{x})} [\ell(x, z)].$$

Theorem (D-Xiao '20)

Suppose $\rho \in (0, \frac{1}{2})$. Run an online algorithm where in iteration t , nature draws $z_t \sim \mathcal{D}(x_t)$ and declares $\ell_t(x_t) = \ell(x_t, z_t)$. Then

$$\mathbb{E} \left[\varphi \left(\frac{1}{t} \sum_{i=1}^t x_i \right) - \varphi(\bar{x}) \right] \leq \frac{\mathbb{E}[R_t]}{(1 - 2\rho)t},$$

Downside: Requires strong assumptions (bounded domain, Lipschitz loss)

Instead, we analyze algorithms directly.

Assumption: (Finite variance) There is a constant $\sigma > 0$ satisfying

$$\mathbb{E}_{z \sim \mathcal{D}(x)} \|\nabla \ell(x, z) - \nabla f_x(x)\|^2 \leq \sigma^2 \quad \forall x.$$

Proximal stochastic gradient (SG)

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \text{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Proximal stochastic gradient (SG)

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \text{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Theorem (D-Xiao '20, Dünner '20)

If $\rho < \frac{1}{2}$, proximal SG finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O} \left(\kappa \cdot \log \left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon} \right) + \frac{\sigma^2}{\alpha \varepsilon} \right) \text{ samples.}$$

Proximal stochastic gradient (SG)

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \text{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Theorem (D-Xiao '20, Dünner '20)

If $\rho < \frac{1}{2}$, proximal SG finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O} \left(\kappa \cdot \log \left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon} \right) + \frac{\sigma^2}{\alpha \varepsilon} \right) \text{ samples.}$$

If $\rho < 1$, proximal SG finds x with $\|x - \bar{x}\|^2 \leq \varepsilon$ using

$$\mathcal{O} \left(\kappa \cdot \log \left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon} \right) + \frac{\sigma^2}{\alpha^2 \varepsilon} \right) \text{ samples.}$$

Proximal stochastic gradient (SG)

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \text{prox}_{\eta r}(x_t - \eta \nabla \ell(x_t, z_t)) \end{array} \right\}$$

Theorem (D-Xiao '20, Dünner '20)

If $\rho < \frac{1}{2}$, proximal SG finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O} \left(\kappa \cdot \log \left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon} \right) + \frac{\sigma^2}{\alpha \varepsilon} \right) \text{ samples.}$$

If $\rho < 1$, proximal SG finds x with $\|x - \bar{x}\|^2 \leq \varepsilon$ using

$$\mathcal{O} \left(\kappa \cdot \log \left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon} \right) + \frac{\sigma^2}{\alpha^2 \varepsilon} \right) \text{ samples.}$$

Remark:

1. Reduces to classical rate if $\rho = 0$ (Lan '10)
2. Last iterate convergence if $r = \delta_C$ in (Dünner-Perdomo-Zrnic-Hardt '20)

Proof sketch: grad deviation controls bias

$$\text{Recall } f_x(x) = \mathbb{E}_{z \sim \mathcal{D}(x)} \ell(x, z) \quad \implies \quad \text{Bias}(x) = \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\|$$

Proof sketch: grad deviation controls bias

$$\text{Recall } f_x(x) = \mathbb{E}_{z \sim \mathcal{D}(x)} \ell(x, z) \quad \implies \quad \text{Bias}(x) = \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\|$$

$$\begin{aligned} \text{Gradient deviation} \quad &\implies \quad \text{Bias}(x) \leq \beta\gamma \cdot \|x - \bar{x}\| \\ &\implies \quad \langle \nabla f_x(x), x - \bar{x} \rangle \geq [f_{\bar{x}}(x) - f_{\bar{x}}(\bar{x})] + \frac{\alpha(1-2\rho)}{2} \|x - \bar{x}\|^2 \end{aligned}$$

Proof sketch: grad deviation controls bias

$$\text{Recall } f_x(x) = \mathbb{E}_{z \sim \mathcal{D}(x)} \ell(x, z) \quad \implies \quad \text{Bias}(x) = \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\|$$

$$\begin{aligned} \text{Gradient deviation} \quad &\implies \quad \text{Bias}(x) \leq \beta\gamma \cdot \|x - \bar{x}\| \\ &\implies \quad \langle \nabla f_x(x), x - \bar{x} \rangle \geq [f_{\bar{x}}(x) - f_{\bar{x}}(\bar{x})] + \frac{\alpha(1-2\rho)}{2} \|x - \bar{x}\|^2 \end{aligned}$$

Lemma: (One-step progress) It holds:

$$2\eta\mathbb{E}[\varphi(x_{t+1}) - \varphi(\bar{x})] \leq (1 - \hat{\alpha}\eta) \mathbb{E}\|x_t - \bar{x}\|^2 - \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + O(\eta^2),$$

where $\hat{\alpha} \approx \alpha(1 - 2\rho)$.

Proof sketch: grad deviation controls bias

$$\text{Recall } f_x(x) = \mathbb{E}_{z \sim \mathcal{D}(x)} \ell(x, z) \quad \implies \quad \text{Bias}(x) = \|\nabla f_x(x) - \nabla f_{\bar{x}}(x)\|$$

$$\begin{aligned} \text{Gradient deviation} &\implies \text{Bias}(x) \leq \beta\gamma \cdot \|x - \bar{x}\| \\ &\implies \langle \nabla f_x(x), x - \bar{x} \rangle \geq [f_{\bar{x}}(x) - f_{\bar{x}}(\bar{x})] + \frac{\alpha(1-2\rho)}{2} \|x - \bar{x}\|^2 \end{aligned}$$

Lemma: (One-step progress) It holds:

$$2\eta \mathbb{E}[\varphi(x_{t+1}) - \varphi(\bar{x})] \leq (1 - \hat{\alpha}\eta) \mathbb{E}\|x_t - \bar{x}\|^2 - \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + O(\eta^2),$$

where $\hat{\alpha} \approx \alpha(1 - 2\rho)$. Combining with strong convexity get

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \leq (1 - \hat{\alpha}\eta) \mathbb{E}\|x_t - \bar{x}\|^2 + O(\eta^2),$$

where $\hat{\alpha} \approx \alpha(1 - \rho)$.

... the rest is standard

Proximal accelerated stochastic gradient (ASG)

Algorithm: (Kulunchakov-Mairal '19)

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(y_{t-1}) \text{ and set } g_t = \nabla \ell(y_{t-1}, z_t), \\ \text{Set } x_t = \text{prox}_{\eta_t r}(y_{t-1} - \eta g_t), \\ \text{Set } y_t = x_t + \frac{1 - \sqrt{\eta \alpha (1 - 2\rho)}}{1 + \sqrt{\eta \alpha (1 - 2\rho)}} (x_t - x_{t-1}). \end{array} \right\}$$

Remark: first proximal ASG due to Ghadimi-Lan '13

Proximal accelerated stochastic gradient (ASG)

Algorithm: (Kulunchakov-Mairal '19)

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(y_{t-1}) \text{ and set } g_t = \nabla \ell(y_{t-1}, z_t), \\ \text{Set } x_t = \text{prox}_{\eta_t r}(y_{t-1} - \eta g_t), \\ \text{Set } y_t = x_t + \frac{1 - \sqrt{\eta \alpha (1 - 2\rho)}}{1 + \sqrt{\eta \alpha (1 - 2\rho)}}(x_t - x_{t-1}). \end{array} \right\}$$

Remark: first proximal ASG due to Ghadimi-Lan '13

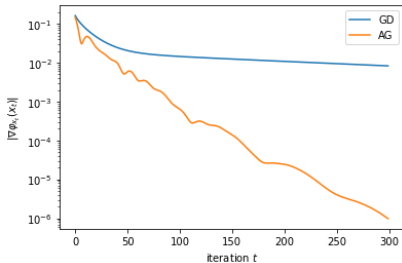
Theorem (D-Xiao '20)

If $\rho \lesssim \kappa^{-1/4}$, proximal ASG finds x satisfying $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

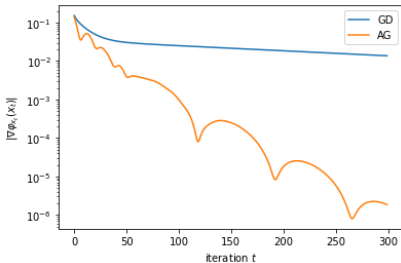
$$\mathcal{O} \left(\sqrt{\kappa} \cdot \log \left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon} \right) + \frac{\sigma^2}{\alpha \varepsilon} \right) \quad \text{samples.}$$

Proof: technical using stoch. estimate sequences (Kulunchakov-Mairal '19)

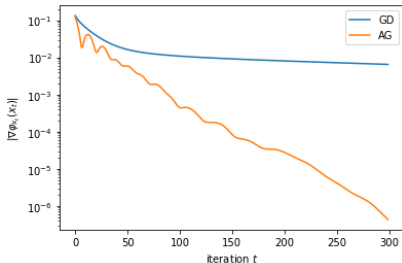
Acceleration works mysteriously well!



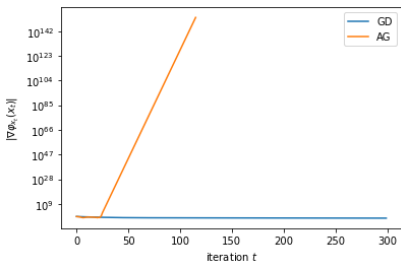
(a) $\gamma = 0$.



(b) $\gamma = 5$.

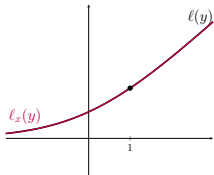


(c) $\gamma = 100$.



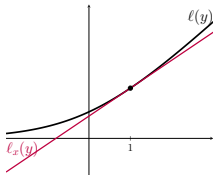
(d) $\gamma = 250$.

Model-based algorithms



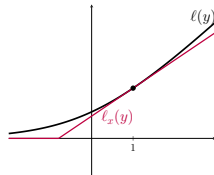
prox-point

$$\ell_x(y) = \ell(y)$$



gradient

$$\ell_x(y) = \ell(x) + \langle \nabla \ell(x), y - x \rangle$$

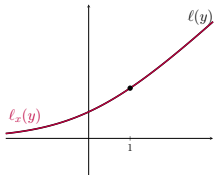


clipped gradient

$$\ell_x(y) = (\ell(x) + \langle \nabla \ell(x), y - x \rangle)^+$$

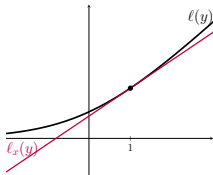
► clipped gradient model introduced in [Asi-Duchi '19](#)

Model-based algorithms



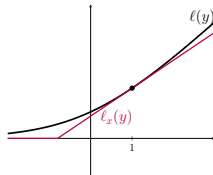
prox-point

$$\ell_x(y) = \ell(y)$$



gradient

$$\ell_x(y) = \ell(x) + \langle \nabla \ell(x), y - x \rangle$$



clipped gradient

$$\ell_x(y) = (\ell(x) + \langle \nabla \ell(x), y - x \rangle)^+$$

► clipped gradient model introduced in [Asi-Duchi '19](#)

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } z_t \sim \mathcal{D}(x_t) \\ \text{Set } x_{t+1} = \underset{y}{\operatorname{argmin}} \ell_{x_t}(y, z_t) + r(y) + \frac{1}{2\eta} \|y - x_t\|^2 \end{array} \right\}$$

Model-based algorithms

Assumption: There exist $\alpha_1, \alpha_2 \geq 0$ such that with $z \sim \mathcal{D}(x)$ have

1. **(Convexity)** $\ell_x(\cdot, z)$ is convex and $\ell_x(\cdot, z) + r$ is α_1 -strongly convex.
2. **(Bias/variance)** It holds:

$$\mathbb{E}_z[\nabla \ell_x(x, z)] = \nabla f_x(x) \quad \text{and} \quad \mathbb{E}_z \|\nabla \ell_x(x, z) - \nabla f_x(x)\|^2 \leq \sigma^2.$$

3. **(Accuracy)** The estimates holds:

$$\mathbb{E}_z[\ell_x(x, z)] = f_x(x) \quad \text{and} \quad \mathbb{E}_z[\ell_x(y, z)] + \frac{\alpha_2}{2} \|x - y\|^2 \leq f_x(y).$$

Model-based algorithms

Assumption: There exist $\alpha_1, \alpha_2 \geq 0$ such that with $z \sim \mathcal{D}(x)$ have

1. **(Convexity)** $\ell_x(\cdot, z)$ is convex and $\ell_x(\cdot, z) + r$ is α_1 -strongly convex.
2. **(Bias/variance)** It holds:

$$\mathbb{E}_z[\nabla \ell_x(x, z)] = \nabla f_x(x) \quad \text{and} \quad \mathbb{E}_z \|\nabla \ell_x(x, z) - \nabla f_x(x)\|^2 \leq \sigma^2.$$

3. **(Accuracy)** The estimates holds:

$$\mathbb{E}_z[\ell_x(x, z)] = f_x(x) \quad \text{and} \quad \mathbb{E}_z[\ell_x(y, z)] + \frac{\alpha_2}{2} \|x - y\|^2 \leq f_x(y).$$

Remark:

- ▶ Similar assumptions in [Davis-Drusvyatskiy '19](#), [Asi-Duchi '19](#)
- ▶ tighter models yield better algorithms [Ryu-Boyd '14](#), [Asi-Duchi '19](#)

Model-based algorithms

Theorem (D-Xiao '20)

If $\frac{\gamma\beta}{\alpha_1 + \alpha_2} < \frac{1}{2}$, algorithm finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O}\left(\frac{L}{\alpha_1 + \alpha_2} \cdot \log\left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon}\right) + \frac{\sigma^2}{(\alpha_1 + \alpha_2)\varepsilon}\right) \quad \text{samples.}$$

Model-based algorithms

Theorem (D-Xiao '20)

If $\frac{\gamma\beta}{\alpha_1 + \alpha_2} < \frac{1}{2}$, algorithm finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O}\left(\frac{L}{\alpha_1 + \alpha_2} \cdot \log\left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon}\right) + \frac{\sigma^2}{(\alpha_1 + \alpha_2)\varepsilon}\right) \quad \text{samples.}$$

If $\frac{\gamma\beta}{\alpha_1 + \alpha_2} < 1$, algorithm finds x with $\|x - \bar{x}\|^2 \leq \varepsilon$ using

$$\mathcal{O}\left(\frac{L}{\alpha_1 + \alpha_2} \cdot \log\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right) + \frac{\sigma^2}{(\alpha_1 + \alpha_2)^2\varepsilon}\right) \quad \text{samples.}$$

Model-based algorithms

Theorem (D-Xiao '20)

If $\frac{\gamma\beta}{\alpha_1 + \alpha_2} < \frac{1}{2}$, algorithm finds x with $\mathbb{E}[\varphi(x) - \varphi(\bar{x})] \leq \varepsilon$ using

$$\mathcal{O}\left(\frac{L}{\alpha_1 + \alpha_2} \cdot \log\left(\frac{\varphi(x_0) - \varphi(\bar{x})}{\varepsilon}\right) + \frac{\sigma^2}{(\alpha_1 + \alpha_2)\varepsilon}\right) \quad \text{samples.}$$

If $\frac{\gamma\beta}{\alpha_1 + \alpha_2} < 1$, algorithm finds x with $\|x - \bar{x}\|^2 \leq \varepsilon$ using

$$\mathcal{O}\left(\frac{L}{\alpha_1 + \alpha_2} \cdot \log\left(\frac{\|x_0 - \bar{x}\|^2}{\varepsilon}\right) + \frac{\sigma^2}{(\alpha_1 + \alpha_2)^2\varepsilon}\right) \quad \text{samples.}$$

Rates for **stochastic proximal point** and **clipped gradient** follow immediately.

Proof sketch: function gap deviation

Lemma: (One-step progress on φ_{x_t}) For every y it holds:

$$2\eta\mathbb{E}[\varphi_{x_t}(x_{t+1})-\varphi_{x_t}(y)] \leq (1-\alpha_2\eta)\mathbb{E}\|x_t-y\|^2 - (1+\alpha_1\eta)\mathbb{E}\|x_{t+1}-y\|^2 + O(\eta^2),$$

Proof sketch: function gap deviation

Lemma: (One-step progress on φ_{x_t}) For every y it holds:

$$2\eta\mathbb{E}[\varphi_{x_t}(x_{t+1}) - \varphi_{x_t}(y)] \leq (1 - \alpha_2\eta)\mathbb{E}\|x_t - y\|^2 - (1 + \alpha_1\eta)\mathbb{E}\|x_{t+1} - y\|^2 + O(\eta^2),$$

Gap deviation $\implies$

$$\begin{aligned}\varphi_{x_t}(x_{t+1}) - \varphi_{x_t}(\bar{x}) &\geq \varphi_{\bar{x}}(x_{t+1}) - \varphi_{\bar{x}}(\bar{x}) - \gamma\beta\|x_{t+1} - \bar{x}\| \cdot \|x_t - \bar{x}\| \\ &\geq \varphi_{\bar{x}}(x_{t+1}) - \varphi_{\bar{x}}(\bar{x}) - \frac{\gamma\beta}{2}\|x_{t+1} - \bar{x}\|^2 - \frac{\gamma\beta}{2}\|x_t - \bar{x}\|^2,\end{aligned}$$

Proof sketch: function gap deviation

Lemma: (One-step progress on φ_{x_t}) For every y it holds:

$$2\eta\mathbb{E}[\varphi_{x_t}(x_{t+1}) - \varphi_{x_t}(y)] \leq (1 - \alpha_2\eta)\mathbb{E}\|x_t - y\|^2 - (1 + \alpha_1\eta)\mathbb{E}\|x_{t+1} - y\|^2 + O(\eta^2),$$

Gap deviation $\implies$

$$\begin{aligned}\varphi_{x_t}(x_{t+1}) - \varphi_{x_t}(\bar{x}) &\geq \varphi_{\bar{x}}(x_{t+1}) - \varphi_{\bar{x}}(\bar{x}) - \gamma\beta\|x_{t+1} - \bar{x}\| \cdot \|x_t - \bar{x}\| \\ &\geq \varphi_{\bar{x}}(x_{t+1}) - \varphi_{\bar{x}}(\bar{x}) - \frac{\gamma\beta}{2}\|x_{t+1} - \bar{x}\|^2 - \frac{\gamma\beta}{2}\|x_t - \bar{x}\|^2,\end{aligned}$$

Combining with Lemma:

$$2\eta\mathbb{E}[\varphi_{\bar{x}}(x_{t+1}) - \varphi_{\bar{x}}(y)] \leq (1 - \hat{\alpha}_2\eta)\mathbb{E}\|x_t - \bar{x}\|^2 - (1 + \hat{\alpha}_1\eta)\mathbb{E}\|x_{t+1} - \bar{x}\|^2 + O(\eta^2),$$

where $\hat{\alpha}_1 = \alpha_1 - \gamma\beta$ and $\hat{\alpha}_2 = \alpha_2 - \gamma\beta$

... the rest is standard

Inexact repeated minimization (IRM)

In typical applications:

“deployment of a learning rule” $\overset{\text{costs}}{\gg}$ “sampling”

Dünner-Perdomo-Zrnic-Hardt '20:

establish “deployments/samples” trade-off for IRM w/ projected SG method

Theorem (D-Xiao '20)

If $\rho < 1$, can implement IRM with all previous algorithms with same sample efficiency and only $\frac{1}{1-\rho} \log(1/\varepsilon)$ deployments.

Details in the paper:

- ▶ “Stochastic optimization with decision-dependent distributions”
D-Xiao (2020), arxiv.org/abs/2011.11173

Main references:

- ▶ “Performative prediction”
Perdomo-Zrnic-Dünner-Hardt (ICML 2020)
- ▶ “Stochastic optimization for performative prediction”
Dünner-Perdomo-Zrnic-Hardt (NeurIPS 2020)
- ▶ “Strategic classification”
Hardt-Megiddo-Papadimitriou-Wootters (ACM ITCS '16)

Details in the paper:

- ▶ “Stochastic optimization with decision-dependent distributions”
D-Xiao (2020), arxiv.org/abs/2011.11173

Main references:

- ▶ “Performative prediction”
Perdomo-Zrnic-Dünner-Hardt (ICML 2020)
- ▶ “Stochastic optimization for performative prediction”
Dünner-Perdomo-Zrnic-Hardt (NeurIPS 2020)
- ▶ “Strategic classification”
Hardt-Megiddo-Papadimitriou-Wootters (ACM ITCS '16)

Thank you!