Mirror Descent and Multi-Level Optimization

Mark Schmidt

UBC

November 2015

Quadratic Approximations

- Recall gradient descent is based on quadratic approximation,

$$x^{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2\alpha_k} \|x - x^k\|^2 \right\}.$$

Quadratic Approximations

- Recall gradient descent is based on quadratic approximation,

$$x^{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2\alpha_k} \|x - x^k\|^2 \right\}.$$

- We've discussed a variety of variations on this:
  - Add extrapolation: momentum/heavy-ball/Nesterov.
  - Replace $\|x - x^k\|^2$ with $\|x - x^k\|_H^2$: Newton.
  - Replace $f'(x^k)$ with $g^k \in \partial f(x^k)$: subgradient.
  - Replace $f'(x^k)$ with $f_i'(x^k)$: stochastic gradient.
  - Replace $f'(x^k)$ with memory of old gradients: SAG.

Quadratic Approximations

- Recall gradient descent is based on quadratic approximation,

$$x^{k+1} = \mathop{\mathrm{argmin}}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2\alpha_k} \|x - x^k\|^2 \right\}.$$

- We've discussed a variety of variations on this:
  - Add extrapolation: momentum/heavy-ball/Nesterov.
  - Replace $\|x - x^k\|^2$ with $\|x - x^k\|_H^2$: Newton.
  - Replace $f'(x^k)$ with $g^k \in \partial f(x^k)$: subgradient.
  - Replace $f'(x^k)$ with $f_i'(x^k)$: stochastic gradient.
  - Replace $f'(x^k)$ with memory of old gradients: SAG.
  - Replace $\mathbb{R}^d$ with convex set $\mathcal{C}$: projected gradient.
  - Add extra non-smooth term $g(x)$: proximal-gradient.
  - Adding more terms and a λ update: ADMM.
  - Use compact $\mathcal{C}$ and remove $\|x - x^k\|^2$ term: Frank-Wolfe.
  - Replace $f'(x^k)$ with $f_j'(x^k)e_j$: coordinate descent.

Quadratic Approximations

- Recall gradient descent is based on quadratic approximation,

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{2\alpha_k}\|x - x^k\|^2 \right\}.$$

- We've discussed a variety of variations on this:
  - Add extrapolation: momentum/heavy-ball/Nesterov.
  - Replace $\|x - x^k\|^2$ with $\|x - x^k\|_H^2$: Newton.
  - Replace $f'(x^k)$ with $g^k \in \partial f(x^k)$: subgradient.
  - Replace $f'(x^k)$ with $f_i'(x^k)$: stochastic gradient.
  - Replace $f'(x^k)$ with memory of old gradients: SAG.
  - Replace $\mathbb{R}^d$ with convex set $\mathcal{C}$: projected gradient.
  - Add extra non-smooth term $g(x)$: proximal-gradient.
  - Adding more terms and a λ update: ADMM.
  - Use compact $\mathcal{C}$ and remove $\|x - x^k\|^2$ term: Frank-Wolfe.
  - Replace $f'(x^k)$ with $f_j'(x^k)e_j$: coordinate descent.
- You can mix/match: proximal quasi-Newton methods, block-coordinate Frank-Wolfe, proximal-SVRG, etc.
- Today: algorithms based on non-quadratic approximations.

Non-Quadratic Approach 1: Mirror Descent

- Modern view of mirror descent iteration:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{\alpha_k} D(x, x^k) \right\},$$

  where $D(x, x^k)$ is a Bregman divergence (BD).
- Informally: BDs are functions that act like $\|x - x^k\|^2$.

Non-Quadratic Approach 1: Mirror Descent

- Modern view of mirror descent iteration:

$$x^{k+1} = \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{\alpha_k} D(x, x^k) \right\},$$

  where $D(x, x^k)$ is a Bregman divergence (BD).
- Informally: BDs are functions that act like $\|x - x^k\|^2$.
- Formally, given a strictly-convex function h, BD is defined by

$$D_h(y, x) = h(y) - h(x) - \langle h'(x), y - x \rangle,$$

  difference between $h(y)$ and first-order Taylor expansion at x.

Non-Quadratic Approach 1: Mirror Descent

- Modern view of mirror descent iteration:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{\alpha_k} D(x, x^k) \right\},$$

  where $D(x, x^k)$ is a Bregman divergence (BD).
- Informally: BDs are functions that act like $\|x - x^k\|^2$.
- Formally, given a strictly-convex function h, BD is defined by

$$D_h(y, x) = h(y) - h(x) - \langle h'(x), y - x \rangle,$$

  difference between $h(y)$ and first-order Taylor expansion at x.
- Properties:
  - Non-negative: $D_h(y, x) \geq 0$.
  - Strictly convex in y (though not necessarily in x).
  - BD of convex conjugate: $D_{h^*}(f'(y), f'(x)) = D_h(x, y)$.

Examples of Bregman Divergences

- Deifnition of Bregman divergence for strongly-convex h,

$$D_h(y, x) = h(y) - h(x) - \langle h'(x), y - x \rangle.$$

- For $h(x) = \|x\|^2$, we get $D_h(y, x) = \|y - x\|^2$:

Examples of Bregman Divergences

- Deifnition of Bregman divergence for strongly-convex h,

$$D_h(y, x) = h(y) - h(x) - \langle h'(x), y - x \rangle.$$

- For $h(x) = \|x\|^2$, we get $D_h(y, x) = \|y - x\|^2$:

$$\begin{aligned} D_h(y, x) &= \|y\|^2 - \|x\|^2 - \langle 2x, y - x \rangle \\ &= \|y\|^2 - \|x\|^2 - \langle 2x, y - x \rangle \pm \langle 2y, y - x \rangle \\ &= \|y\|^2 + \|x\|^2 - 2y^T x = \|y - x\|^2. \end{aligned}$$

- For $h(x) = \|x\|_H^2$, we get $D_h(y, x) = \|y - x\|_H^2$.

Examples of Bregman Divergences

- Deifnition of Bregman divergence for strongly-convex h,

$$D_h(y,x) = h(y) - h(x) - \langle h'(x), y - x \rangle.$$

- If domain is probabilities and h is entropy, $h(x) = \sum_i x_i \log x_i$,

Examples of Bregman Divergences

- Deifnition of Bregman divergence for strongly-convex h,

$$D_h(y,x) = h(y) - h(x) - \langle h'(x), y - x\rangle.$$

- If domain is probabilities and h is entropy, $h(x) = \sum_i x_i \log x_i$,

$$\begin{aligned}
D_h(y,x) &= \sum_i y_i \log y_i - \sum_i x_i \log x_i - \sum_i (1 + \log(x_i))(y_i - x_i) \\
&= \sum_i y_i \log y_i - \sum_i x_i \log x_i - \sum_i y_i + \sum_i x_i - \sum_i (y_i - x_i \\
&= \sum_i y_i \log y_i - \sum_i x_i \log x_i - \sum_i (y_i - x_i) \log x_i. \\
&= \sum_i y_i \log y_i - \sum_i x_i \log x_i - \sum_i y_i \log x_i + \sum_i x_i \log x_i. \\
&= \sum_i y_i \log y_i - \sum_i y_i \log x_i = \sum_i y_i \log \frac{y_i}{x_i} \triangleq D_{KL}(y||x).
\end{aligned}$$

which is the Kullback-Leibler divergence.

Entropic Descent and Exponentiated Gradient

- Consider optimizing over the probability simplex

$$\operatorname*{argmin}_{x \geq 0, \sum_i x_i = 1} f(x).$$

- Consider using mirror descent with the KL divergence,

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{\alpha_k} D_{KL}(x || x^k) \right\}.$$

Entropic Descent and Exponentiated Gradient

- Consider optimizing over the probability simplex

$$\underset{x \geq 0, \sum_i x_i = 1}{\operatorname{argmin}} \quad f(x).$$

- Consider using mirror descent with the KL divergence,

$$x^{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x^k) + \langle f'(x^k), x - x^k \rangle + \frac{1}{\alpha_k} D_{KL}(x || x^k) \right\}.$$

- The update for each variable j is given by

$$x_j^{k+1} = \frac{x_j^k \exp(-\alpha_k f_j'(x^k))}{\sum_{j'} x_{j'}^k \exp(-\alpha_k f_{j'}'(x^k)}.$$

- If x^0 satisfies constraints, all iterations satisfy constraints.
- Called entropic descent or exponentiated gradient.

Convergence Rate of Exponentiated Gradient

- Regular projected sub-gradient has a rate of $O(1/\sqrt{k})$.
  - Constant has no dependence on n.
  - Constants depends on Lipschitz constant in ℓ_2-norm, L_2.

Convergence Rate of Exponentiated Gradient

- Regular projected sub-gradient has a rate of $O(1/\sqrt{k})$.
  - Constant has no dependence on n.
  - Constants depends on Lipschitz constant in ℓ_2-norm, L_2.
- Projected sub-gradient mirror descent also has $O(1/\sqrt{k})$.
  - Constant has a $\log(n)$ dependence.
  - Constant depends on Lipschitz in ℓ_1-norm, L_1.

Convergence Rate of Exponentiated Gradient

- Regular projected sub-gradient has a rate of $O(1/\sqrt{k})$.
  - Constant has no dependence on n.
  - Constants depends on Lipschitz constant in ℓ_2-norm, L_2.
- Projected sub-gradient mirror descent also has $O(1/\sqrt{k})$.
  - Constant has a $\log(n)$ dependence.
  - Constant depends on Lipschitz in ℓ_1-norm, L_1.
- We have $L_1 \leq L_2 \leq \sqrt{n}L_1$:
  - If left is tight, mirror descent is worse by $\sqrt{\log(n)}$.
  - In right is tight, mirror descent improves $\sqrt{n}$ to $\sqrt{\log(n)/n}$.

Convergence Rate of Exponentiated Gradient

- Strongly-convex: rate improves to $O(\log(t)/t)$.
- Stochastic mirror descent: rates stay the same.
- Smooth case: accelerated $O(1/t^2)$ variants.

Convergence Rate of Exponentiated Gradient

- Strongly-convex: rate improves to $O(\log(t)/t)$.
- Stochastic mirror descent: rates stay the same.
- Smooth case: accelerated $O(1/t^2)$ variants.
- Learning theory [Kivinen & Warmuth, 1997]:
  - Exponentiated gradient is better if few relevant variables.
- Pre-SAG: For log-linear models, dual block exponentiated gradient has linear rate [Collins et al., 2007].

Non-Quadratic Approach 2: Multi-Level Methods

- We want to minimize a smooth function F,

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, F(x),$$

  and it is very expensive to evaluate F.
- But we quickly optimize a related cheap function f.

Non-Quadratic Approach 2: Multi-Level Methods

- We want to minimize a smooth function F,

$$\underset{x \in \mathbb{R}^d}{\operatorname{argmin}} F(x),$$

  and it is very expensive to evaluate F.

- But we quickly optimize a related cheap function f.
- Examples:
  - Total-variation on a big image (F) or smaller version (f).
  - Fitting CRF with variational (F) or pseudolikelihood (f).
  - Fitting model on full data (F) or small sub-samples (f).
  - Differential equation on fine grid (F) vs. coarse grid (f).
- Could have more than 2 levels, but we'll focus on 2.

Multi-Level Optimization

- Multi-level optimization methods repeat three steps:
  1. Cheap minimization of modified f (can start with $v_0 = 0$).

$$y^k = \underset{x \in \mathbb{R}^d}{\text{argmin}}\, f(x) + \langle v_k, x \rangle.$$

Multi-Level Optimization

- Multi-level optimization methods repeat three steps:
  1. Cheap minimization of modified f (can start with $v_0 = 0$).
  $$y^k = \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) + \langle v_k, x \rangle.$$
  2. Use y^k to give descent direction,
  $$x^{k+1} = x^k - \alpha_k (x^k - y^k).$$

Multi-Level Optimization

- Multi-level optimization methods repeat three steps:
  1. Cheap minimization of modified f (can start with $v_0 = 0$).
  $$y^k = \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) + \langle v_k, x \rangle.$$
  2. Use y^k to give descent direction,
  $$x^{k+1} = x^k - \alpha_k (x^k - y^k).$$
  3. Set v_k to satisfy first-order coherence:
  $$v_{k+1} = \frac{L_f}{L_F} F'(x^{k+1}) - f'(x^{k+1}).$$
- Above we assume that F and f have same parameters:
  - Add projection if defined on different variables.
  - Called 'restriction' and 'prolongation'.
- Linear rate depending on various factors [Parpas et al., 2014].

First-Order Coherence Condition

- Consider the first iteration of gradient descent on f,

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} f(x^0) + \langle f'(x^0), x - x^0 \rangle + \frac{L_f}{2}\|x - x^0\|^2.$$

- Makes progress on f, but no relation to F.

First-Order Coherence Condition

- Consider the first iteration of gradient descent on f,

$$\underset{x \in \mathbb{R}^d}{\text{argmin}} \; f(x^0) + \langle f'(x^0), x - x^0 \rangle + \frac{L_f}{2}\|x - x^0\|^2.$$

- Makes progress on f, but no relation to F.
- Now consider the modified function

$$h(x) = f(x) + \langle v_k, x \rangle = f(x) + \langle \frac{L_f}{L_F} F'(x^0) - f'(x^0), x \rangle$$
$$h'(x) = f'(x) + \frac{L_f}{L_F} F'(x^0) - f'(x^0).$$

First-Order Coherence Condition

- Consider the first iteration of gradient descent on f,

$$\underset{x \in \mathbb{R}^d}{\text{argmin}} \; f(x^0) + \langle f'(x^0), x - x^0 \rangle + \frac{L_f}{2}\|x - x^0\|^2.$$

- Makes progress on f, but no relation to F.
- Now consider the modified function

$$h(x) = f(x) + \langle v_k, x \rangle = f(x) + \langle \frac{L_f}{L_F} F'(x^0) - f'(x^0), x \rangle$$

$$h'(x) = f'(x) + \frac{L_f}{L_F} F'(x^0) - f'(x^0).$$

- By playing with argmins, first iteration on h gives

$$\underset{x \in \mathbb{R}^d}{\text{argmin}} \; F(x^0) + \langle F'(x^0), x - x^0 \rangle + \frac{L_F}{2}\|x - x^0\|^2,$$

which is gradient descent on F.

- But could make progress if F and f.

Non-Quadratic Approaches 3+4

- Mairal [2013,2014] considers general surrogate optimization:

$$x^{t+1} = \underset{x \in \mathcal{C}}{\operatorname{argmin}} \left\{ f(y) \right\},$$

- Cheap function f upper bounds expensive function F.
- Function values and gradients of f and F agree at x^t.
- Function $f' - F'$ is Lipschitz-continuous.
- Obtains $O(1/k)$ and linear rates depending on $f - F$.
- Hennig & Kiefel [2013] propose non-parametric quasi-Newton:
  - View quasi-Newton methods as MAP estimators.
  - New method incorporates all previous gradients.

Summary

- Mirror descent considers other Bregman divergences.
  - Advantages for optimization over simplex.
  - Other interesting divergences/problems?
- Multi-level/surrogate consider cheap f and expensive F.
  - Great for problems that have multiple resolutions.
  - Useful for ML methods like graphical models?
- Room for improvement over classic quadratic approximations:
  - Non-parametric quasi-Newton.